Proc. Natl. Acad. Sci. USA
Vol. 94, p. 8923, August 1997
From the Academy

Discussion of evaluation of federal laboratories*

NORMAN METZGER

Executive Director, Commission on Physical Sciences, Mathematics, and Applications, National Research Council, Washington, DC 20418

Federal agencies are now beginning work on their fiscal year 1999 budget requests. Compounding the usual difficulties in budgeting is that *all* agencies must for the first time justify their requests in terms of outcomes to be achieved, using pre-approved measures. That requirement for results-oriented management is mandated by the Government Performance and Results Act of 1993 (GPRA). Pilot tests of the GPRA requirements—specifically, measures that agencies can use to assess outcomes—have been under way for the past few years at several agencies, including two that support scientific research, the National Science Foundation (NSF) and the Army Research Laboratory (ARL).

The conundrum for the scientific community is obvious. It is not hard to imagine outcome measures for processing of Social Security forms, procuring new weapons, or enforcing environmental standards. However, what outcome measures should the government apply to the use of public funds to support fundamental research? Managers at NSF, ARL, and other agencies are now wrestling with that question. It is not simple. As the 1995 NAS/NAE/IOM report on *Allocation of Federal Funds for Science and Technology* observed:

> Any system to allocate resources should be guided by explicit goals, expressing the underlying philosophy and criteria for evaluating performance. But a clear message emerges from the abundant recent writing on applying performance measures to research and development: it is a complicated business. The science of metrics documents that most measures are incomplete, and mindless application actually can undermine the very functions such measures are intended to improve. Just as the tyranny of quarterly bottom lines can frustrate long-term corporate planning, so also can science be distorted by simple indicators such as publication counts, citation counts, patent counts, doctorates produced, or user satisfaction ratings. These are useful, but incomplete, measures. (p. 27)

These comments are apt, but the law is in place and the issue for the federal agencies that serve as stewards for the nation's scientific enterprise is not whether to comply but how. One response to the "how" is suggested by evaluation work conducted by the National Research Council. The NRC has at times evaluated federal scientific and technical programs, such as those of the Air Force Office of Scientific Research, the Office of Naval Research, and others. Especially worth noting are NRC evaluations of two major federal laboratories: the intramural programs of the National Institute of Standards and Technology (NIST) and of the ARL because, like the GPRA requirements, they occur annually. The NIST evaluations are well seasoned, because they have been done annually by the NRC since 1959; those for the ARL are very new—the first review was finished and transmitted to the ARL in December.

These are substantial programs: the fiscal year 1997 budget levels for the NIST laboratories and ARL are, respectively, $268 million and $393 million, and have personnel levels of 3,000 and 2,500 each. The general approach for both reviews is the same: a board conducts and monitors the review with the aid of panels it forms to review specific laboratories, in the case of NIST, and mission areas, in the case of ARL. The NIST Board has panels on—to pick a few arbitrarily—physics, manufacturing engineering, and building and fire research; and the ARL Board has panels on vehicle technology, weapons and material research, and human research and engineering, among others.

Panels meet once for two to three days on site, with part of that time dedicated to briefings, tours, and demonstrations on work under way and the remaining time to the panels meeting in executive sessions, when they begin drafting their reports, iterating and finishing up through the usual means—e-mail, faxes, etc. The panel chairs meet subsequently with their parent boards to summarize their findings and to enable the board to agree on issues common across the laboratories being evaluated. The boards then publish their own reports, with those of its panels appended.

That is the process. What is the value? What are the strengths? How can they become even more effective?

NIST directors have repeatedly affirmed the importance of the evaluation by the NRC of their laboratories; and, indeed, one can point to many changes that have followed NRC judgments on the quality and relative importance of specific programs. It is too early to tell with the ARI. The report is critical of several parts of the ARL program, and the real impact will be seen in the responses of ARL management and the higher echelons of the Army to the criticisms.

The strength is clear. Federal laboratories are provided with independent peer reviews of their programs, with the NRC serving to validate that the right range of expertise is fitted to the laboratory or mission being evaluated, that the judging is done by a committee of some of the country's best on the topic, and, through the validation of panel judgments by parent boards and then by the NRC review process, that the judgments are fair and balanced.

Whether these two boards will serve as a template for other agencies in evaluating their own laboratories is too early to tell. However, it would not be surprising if the examples set by these two boards becomes a heuristic for other agencies as they seek effective ways to respond to demands of GPRA that are not antithetical to fundamental canons of science for judging merit. To reify that, the ARL, in addition to the outside peer review provided by the NRC, also uses internal metrics to assess its management. As the value of independent peer assessments to judging the value of federal investments becomes clear, they may become a standard part of GPRA compliance across the government. If this happens, then GRPA, rather than dampening R&D innovation as might be expected if it required rote compliance with objective goals, may actually provide a net gain, by ensuring that even those government labs that had been insular, such as the ARL, begin measuring themselves against the quality standards of the broader R&D community.

*This paper is part of the fifth installment of the new feature, "From the Academy." The first installment appeared in the March 4, 1997 issue. "From the Academy" will be presented occasionally as new NRC reports appear and as essays on the NAS are prepared.

0027-8424/97/948923-1$0.00/0
PNAS is available online at http://www.pnas.org.